Maximizing Repository Exposure

Conference Proposal
Services, Interfaces, Scholarly Communication

Introduction

Enhancing online visibility for an institution or its (scientific) digital output, is one of the major goals in many repository projects. This proposal attempts to identify different repository properties, that have an influence on successful online exposure, tries to measure these properties and provides recommendations on optimizing these properties.

Measuring successful online exposure

In this proposal, the assumption is made that repository usage statistics provide a useful RELATIVE indicator of the repository's effectiveness in exposing its contents and generating visibility for the institution.

As a result, properties that are assumed to have positive influence on the online exposure, should have a positive influence on repository usage statistics such as the number of (unique) visits, bitstream downloads or other indicators.

Following properties have been selected for further investigation of their effect on online repository exposure:

Language - When the repository interface and it's contents are presented in a language with a wide geographical coverage, such as English and Spanish, one can expect a more geographically distributed audience of visitors, and possibly a broader impact.

Accessibility through search engines - With search engines such as Google continually exploring and indexing new content on the internet, they have become the entry point for web-browse actions for many users. This proposal attempts to discover how important search engines are for repositories, and looks at optimization.

Content Quantity - Adding more content to the repository could be a successful strategy to attract more visitors. This proposal aims to verify whether the number of items in the repositories correlates with the number of visitors. Also the impact of full-text presence is interesting to explore.

Bram Luyten

@mire nv
Romeinse straat 18
3001 Heverlee

BTW BE 0886.066.294
+32 2 888 29 56
info@mire.be
www.atmire.com

Comparing repository usage

One could argue that the approach of comparing repository usage statistics doesn't take into account indirect exposure generation that happens, for example, when an external harvesting application harvests all of the repositories contents, and continues to serve these contents online. It's true that, although the initial harvesting effort can be included in the repository visit statistics, that additional exposure, generated by the harvester is not taken into account.

In reality, most harvesting platforms today still do not harvest the full-text from a repository, but offer a link to the original repository where a user can download the full-text. Because this approach generates additional visits and downloads for the original repository, these effects ARE taken into account in this research.

To ensure a commonly shared standard in the logging of online visits to the repository, only repositories who make use of Google Analytics, for logging the visits, were included in this research project. Because of its wide adoption, its single and straightforward way to integrate this in the repository, **Google Analytics provides a real common ground of comparison for numbers of online visits**.

This research does not make the claim that these numbers of online visits are correct in an absolute way, but it does rely on the fact that these numbers provide a reliable relative indicator, in order to form propositions about evolutions of visits over time.

Included institutions and their repositories

Following institutions and their staff kindly provided access to their statistics and their appreciated collaboration. As this research is ongoing, this list continues to grow.

National Library of Finland - DORIA	http://oa.doria.fi
University of Göteborg - GUPEA	http://gupea.ub.gu.se
University of Helsinki - E-Theses	http://ethesis.helsinki.fi/en
Malmö University - MUEP	http://mah.se/muep
Katholieke Universiteit Leuven - LIRIAS	http://lirias.kuleuven.be
Texas Digital Library - TDL	http://repositories.tdl.org/tdl
Oregon State University - ScholarsArchive	http://ir.library.oregonstate.edu
University of Auckland - Researchspace	http://researchspace.auckland.ac.nz
University of Pardubice – Digital Library	http://dspace.upce.cz:8443
Lessius Hogeschool - LIRIAS	http://lirias.lessius.eu
Hogeschool Universiteit Brussel - LIRIAS	http://lirias.hubrussel.be
Høgskolen i Telemark - TEORA	http://teora.hit.no/dspace/

Bram Luyten

@mire nv
Romeinse straat 18
3001 Heverlee

BTW BE 0886.066.294
+32 2 888 29 56
info@mire.be
www.atmire.com

Aberystwyth University - CADAIR http://cadair.aber.ac.uk/

All of these institutions have included the Google Analytics tracker code, into the footer of every page in the repository and have at least data over a period of six months time.

Research Questions and Preliminary Findings

Here is an (non-conclusive) overview of the specific research questions, attempted approaches to tackle those questions and some preliminary findings. As the number of participating repositories grows, and more people get involved in the research, more solid conclusions should be available by the time of the conference.

Do harvesting initiatives impact the exposure for repositories and content ?

Evidence was found that for 2 repositories in the sample, a national harvesting initiative consistently generated over 30% of incoming traffic.

Is there a correlation between the number of items and the number of visits ?

Correlation (pearson coefficient, based on linear regression) was calculated between the total number of visits in january 2009, and the number of items in the repository in january.

Initially, when looking at the whole sample, the coefficient of **0.211** wasn't very encouraging. However, when looking at the scatter plot we saw a very obvious outlier at the lower right: a repository featuring more than 150.000 items, but "only" around 28.000 visits. While the other repositories have minimum of 25% items containing at least one bitstream (full-text), this repository only features 3.5% full-text items.

This example illustrates that a mere high number of items doesn't guarantee the same number of monthly visits per item, compared to repositories containing more full text. However, because this is the sole example in this research, it doesn't prove that a high percentage of full-text items guarantees a steady number of monthly visits per item.

When we take this exceptional repository out of the sample, we are getting following scatter plot, and a dramatic improvement to **0.82**

What would be the mean number of visits per item ?

We have found a basis to believe that the number of visits is correlated with the number of items in the repository. As a consequence, it's reasonable to investigate whether the

Bram Luyten

@mire nv
Romeinse straat 18
3001 Heverlee

BTW BE 0886.066.294
+32 2 888 29 56
info@mire.be
www.atmire.com

mean number of visits per item could be a useful indicator to determine how well your repository content is performing in terms of generating exposure.

When taking into account all the repositories, we measured an average of **3.08 monthly visits** per item. The visits per item for all the repositories are illustrated in this graph:

We notice two interesting extreme outliers. They are the main reason why the standard deviation is **3.5** here (which would not be a good case for a reliable mean). The repository we discussed in the previous example, only had 0.16 visitors per item. The other outlying repository had an amazing 12 visits per item. At this point, it remains unclear how the repository is able to generate 48 280 unique monthly visits, with "only" 3783 items. No irregularity has been detected in the use of the tracker but investigation continues.

When discarding those two averages, we get an average of **2,4** visits per item and a much more acceptable standard deviation of **1.76**.

Following averages over the 15 included repositories will be examined for relevance, especially with respect to the aforementioned outlying repository.

Difference between number of visits in december 2008 and january of 2009
+21%

Average number of items
18942

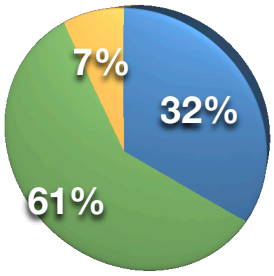
Average percentage of items containing at least one bitstream (calculated over 11 repo's)
27%

Average percentage of national visits in december 2008
56%

Average percentage of visits generated by referring sites
32.1%

Average percentage of visits generated by search engines
60.5%

Average percentage of visits generated by direct traffic
7,3%



We aim to further continue this research and illustrate these visible trends in the ratio between traffic originating from referring sites, direct traffic and search engines.

Bram Luyten

@mire nv
Romeinse straat 18
3001 Heverlee

BTW BE 0886.066.294
+32 2 888 29 56
info@mire.be
www.atmire.com